ISSN 2007-9737

Enhancing the Detection of Sexist Messages through a Multi-Profile-Based Ensemble Approach

Martha Paola Jimenez-Martinez[1], Irvin Hussein Lopez-Nava[1], Manuel Montes-y-Gómez[2,*]

[1] Centro de Investigación Científica y de Educación Superior de Ensenada, Mexico

[2] Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico

{jimenezmp, hussein}@cicese.edu.mx, mmontesg@inaoep.mx

Abstract. Sexism in language perpetuates harmful stereotypes, especially in cultures with deeply ingrained traditional gender roles, such as Mexico. While detection of misogynistic content in English has advanced, detecting sexist language in Spanish is less explored. This study uses the EXIST corpus, annotated by various demographic groups, to examine differing perceptions of sexism across genders and ages. Our analysis finds significant perception discrepancies, with 25% of texts showing disagreements between male and female annotators. We propose an ensemble classification model that integrates outputs from gender-specific and age-specific models based on ROBERTuito, achieving an F1 score of 0.854. To gain insights into our best classifier's decision-making, we present an error analysis based on the visualization of attention weights, which helps us identify the most relevant words in the detection of subtle sexism. Additionally, we leverage ChatGPT's capabilities to model language nuances, generating potential interpretations of texts associated with the classifications provided by our approach. This study underscores the importance of demographic considerations in sexist language detection and demonstrates that combining diverse perspectives with advanced techniques can enhance detection in Spanish social media.

Keywords. Sexism, hierarchical attention networks, transformers, social media, ensemble classification, sexism detection.

1 Introduction

Sexism in language refers to the use of expressions that privilege one gender over another, perpetuating stereotypes and prejudices that can be particularly harmful to women [19]. This type of discrimination is rooted in biological differences and is manifested through attitudes, biases, and stereotypes that imply the inferiority of one gender compared to another [21].

People born or living in Mexico are aware that our culture, like many Latin cultures, is heavily influenced by "machismo"[1]. It is common to encounter numerous sexist expressions prevalent in the Spanish language, such as "*Corres como niña*" (You run like a girl) or "*Los hombres no lloran*" (Men don't cry). Men can also be victims, facing expressions that "test" their masculinity. Additionally, women frequently hear phrases like "*Calladita te ves mas bonita*" (You look prettier when you're quiet), aimed at minimizing and silencing them [6].

Research has shown that regions with higher rates of misogynistic tweets also have higher rates of domestic and family violence [3]. Another study found a correlation between the increase in misogynistic language on the X platform and higher real-life rates of sexual violence [9]. These findings highlight the importance of comprehensive research from various perspectives, including computational methods for the automatic identification of such harmful content. Integrating these methods is crucial in addressing and mitigating the

[1] "Male behaviour that is strong and forceful, and shows very traditional ideas about how men and women should behave" [4]

impact of online misogynistic language on offline gender-based violence incidents.

Hate speech is widely recognized as a complex issue, even among experts familiar with its various definitions [14]. Sexism, a specific form of hate speech, is similarly challenging to define due to its nuanced implications across different contexts. For instance, the Cambridge Dictionary defines sexism as "(actions based on) the belief that the members of one sex are less intelligent, able, skilful, etc. than the members of the other sex, especially that women are less able than men" [5]. This definition tends to focus on discrimination against women, highlighting the historical and social contexts where sexism has disproportionately affected them.

However, it implicitly allows for the application to any gender, although this is not emphasized. On the other hand, the Royal Spanish Academy defines sexism more broadly as "discrimination against individuals based on their sex" (in Spanish: *discriminación de las personas por razón de sexo*) [15]. This definition is more inclusive, leaving the term "sex" open to interpretation, potentially encompassing non-binary individuals and the broader LGBTQ+ community. By doing so, it acknowledges the evolving understanding of gender and the different forms of discrimination that can occur beyond the binary conception of sex. These varying definitions underscore the complexity of defining sexism, as each reflects different cultural and linguistic nuances.

For instance, the statement "She's good at math for a woman" can appear as a compliment on the surface but carries a sexist implication that women are generally not expected to excel in math. Traditional methods might miss this underlying bias due to the phrase's seemingly positive tone, whereas recognizing the sexism requires understanding the context and the implied stereotype. In contrast, Transformer-based models leverage the context of the text to address some of these challenges by capturing the nuances and subtleties of language. However, even these advanced models face difficulties in accurately perceiving sexism across different age groups and genders, highlighting the importance of ensemble methods that combine diverse perspectives for improved detection.

Several notable workshops have significantly advanced the field of hate speech detection, including The Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA) [2]; The Workshop on Online Abuse and Harms (WOAH) [11]; and The International Workshop on Semantic Evaluation (SemEval) [12]. In the latter, a task on Explainable Detection of Online Sexism (EDOS) was introduced, where approximately 90% of the participants, across various subtasks, chose transformer-based architectures. Over the years, various approaches have been employed starting with lexicon-based methods and traditional machine learning models such as SVMs.

As the field advanced, deep learning techniques, including RNNs and CNNs, became popular, followed by embedding-based models and fine-tuning strategies with BERT-like architectures. Recently, transformer-based models like RoBERTa, DeBERTa, and BERT have dominated the field, with some approaches also utilizing large language models like GPT-2, GPT-3, PaLM, and OPT, reflecting the evolution from simpler models to complex, context-aware systems.

Despite the progress in detecting sexist in English tweets, there is a significant gap in the automatic detection of sexist tweets in Spanish [16]. A major effort in this area is the recent organization of the EXIST (sEXism Identification in Social neTworks) shared task [17, 18, 13]. In 2021, with a focus on two primary tasks (Identification and Categorization), approaches for detecting sexism in Spanish utilized a pre-trained multilingual BERT model as well as monolingual BERTmodels [7].

In 2022, also focusing on these two tasks, a novel bi-ensemble approach based on RoBERTa and BERT was introduced, combining transformers pre-trained in both Spanish and English [22]. By 2023, the scope expanded to include three tasks (Identification, Intention, and Categorization).

A cascaded system was developed using GPT-NeoX and BERTIN-GPT-J-6B [20]. Despite the promising results reported by the EXIST task, there is a lack of analysis regarding differences in

the perception of sexism among different groups of people. The EXIST corpus stands out by making the labels of all annotators available, enabling such analysis. However, most studies using this corpus have focused on developing new methods for the automatic detection of sexism without deeply analyzing the differences in annotators' labels. Our primary contribution, which was introduced in [8], is a detailed analysis of labels from six annotator profiles in the EXIST corpus, examining agreement levels and identifying themes causing discordant perceptions.

Our second contribution, presented in this expanded version, involves automatically identifying sexist comments. We propose a novel method for detecting sexist comments that leverages ensemble models to capture the diverse perspectives of different annotator profiles. Specifically, we construct classifiers that represent the views of distinct demographics, including women, men, young adults (18-22), middle-aged adults (23-45), and older adults (46+). Combining these varied opinions through an ensemble approach, we hypothesize that this method will outperform traditional models that rely on consensus-based labeling (hard labels[2]).

Furthermore, we present an error analysis by examining attention values on misclassified instances (false positives and false negatives), and generate explanations using GPT-based prompts, where a tweet and the model's predicted label are input to explore the reasoning behind its decision. Although the models can predict labels, they cannot often explain why. This analysis aids in enhancing the explainability of deep learning approaches, which generally struggle to provide transparent justifications for their predictions.

2 EXIST Corpus

For our analysis and experiments, we utilized the dataset from the sEXism Identification in Social neTworks task at CLEF 2023 [13]. This corpus includes 10,000 entries in both English and Spanish [10]. From the 4,209 labeled instances in Spanish, 3,660 were used for training, with 20% reserved for validation, and 549 instances were used for testing.

Table 1. An example tweet on which annotators reached unanimous agreement

Tweet	eres una golfa, te has pasado por la piedra a medio pepé para llegar al carguito de mierda que tienes, así de triste eres, no vales ni pa caldo fea https://t.co/zEJhy1mKnS *(in english: you're a slut, you've screwed half of the PP to get the crappy position you have, that's how sad you are, you're not even worth broth ugly https://t.co/zEJhy1mKnS)*
Gender	`"F", "F", "F", "M", "M", "M"`
Age	`"18-22", "23-45", "46+", "46+", "23-45", "18-22"`
Labels	`"YES", "YES", "YES", "YES", "YES", "YES"`
Consensus	`"YES"`

The EXIST tasks consist of three main objectives. The first task involves binary classification to determine if a tweet is sexist or not. The second task classifies the message based on three types of author intentions: Direct, Reported, and Judgemental. The third task categorizes the tweet into one or more of five categories: Ideological and Inequality; Role Stereotyping and Dominance; Objectification; Sexual Violence; and Misogyny and Non-Sexual Violence. Our primary focus is on the first task [13].

This corpus differs from typical datasets as it includes labels from six annotator profiles instead of a single definitive label. With 725 annotators, the profiles consist of three women and three men from three age groups: 18-22, 23-45, and 46+. Each text is associated with six labels.

The "hard label[3]" or consensus label is derived from the agreement of these profiles, whether unanimous or majority-based. For example, in Table 1, all annotators agreed on the label.

Conversely, in Table 2, annotators had differing opinions. In this example, only the male annotator aged 23-45 and the male annotator aged 46+ identify the tweet as sexist. Relying

[2]They are obtained through a majority voting strategy on all individual annotations.

[3]"The class annotated by more than 3 annotators is selected, instances for which there is no majority class are removed from this evaluation scheme." [13]

Table 2. An example tweet on which annotators had varying opinions

Tweet	si estás gorda, los espejos de las tiendas te hacen el doble *(in english: if you're fat, store mirrors make you look twice as big)*
Gender	"F", "F", "F", "M", "M", "M"
Age	"18-22", "23-45", "46+", "46+", "23-45", "18-22"
Labels	"NO", "NO", "NO", "NO", "YES", "YES"
Consensus	"NO"

solely on hard labels loses valuable annotator profile information. Therefore, we examined discrepancies in annotations, focusing on gender and age perspectives before model construction.

3 Sexism Perception by Different Annotator Profiles

As a first step, we conducted a qualitative assessment of agreement and disagreement between men and women based on labeled tweets [8]. Out of 3660 texts, 36% showed agreement between men and women that they are sexist, while 39% agreed that they are not sexist. In 12% of texts, men claimed they were sexist while women denied it, and in 13%, women affirmed they were sexist while men denied it. This discrepancy implies potential variations in perception and sensitivity towards sexist content between genders (totaling 25%).

Notable examples of disagreement occur when women label posts as "sexist" while men label them as "non-sexist," as shown in Table 3. These types of posts usually include comments with a humorous tone, which minimize or dismiss women's opinions. Conversely, instances where men categorize a text as "sexist" while women categorize it as "non-sexist" often involve mockery directed towards men.

Our second analysis focuses on the annotators' age groups, where differences, although more subtle, can also be observed among people aged 18-22 years, 23-45 years, and 46+. A considerable agreement is observed between the 18-22 and 23-45 age groups, with 39% of tweets labeled as sexist and 53% labeled as not sexist; conversely, the total disagreement rate was 8%.

Table 3. An example of tweets on which women and men differ in labeling

Labels	Tweet
female: YES **male:** NO	@gishel_paola @PaveloRockstar @giov_ Las mujeres no deben opinar, no porque varias lo hagan, tu también andes haciéndolo *(in english: women shouldn't have an opinion; just because several do, doesn't mean you should be doing it too)*
female: NO **male:** YES	buena foto teodoro pareces una p*ta en decadencia https://t.co/1LecSxbGKi *(in english: nice photo, Teodoro, you look like a decaying whore https://t.co/1LecSxbGKi)*

Table 4. An example of tweets on which individuals ages 18-22, 23-45, and 46+ differ in labeling

Labels	Tweet
18-22: YES **23-45:** NO	@ParentiSol cheto florida 40 te voy a re acosar ahora. *(in english: @ParentiSol posh Florida 40, I'm going to seriously harass you now.)*
18-22: YES **46+:** NO	@PamelaAlemapq @AustinPalao obvio que sii, es guapisimo, chicas un grito saooooooo!!! *(in english: @PamelaAlemapq @AustinPalao obviously, yes, he's super handsome, girls, a shout out!!!)*
23-45: NO **46+:** YES	@PerroChusko @MineduPeru @congresoperu Que el de todas el feminismo es cáncer. *(in english: @PerroChusko @MineduPeru @congresoperu That feminism is cancer to everyone.)*

Despite disagreement rates being below 10%, it is important to examine instances where these varied age groups hold differing views on the concept of sexism. Table 4 illustrates tweets where age groups diverged in their labels of sexism.

It is important to clarify that this analysis was conducted using a reduced dataset of 1333 instances due to the need to identify and select tweets that appeared in all three age groups. This reduction was necessary because, in these groups, we lost data due to the majority vote on their labels. These disagreements represent a minority within the dataset, indicating that the

perception of sexism is more uniform across different age groups compared to the differences observed between genders.

4 Integrating Age and Gender Perspectives through Ensemble Classification

Clear differences in model performance have been observed when comparing gender groups and age groups [8], which suggests that demographic characteristics can markedly influence the accuracy and effectiveness of classification models. In this context, it has been demonstrated that segmenting annotators into specific groups or profiles can considerably improve classification performance in a dataset of abusive language; rather than training a single comprehensive classifier encompassing all data, the strategy of utilizing multiple classifiers specialized in different demographic profiles has proven to be more effective [1].

This improvement not only underscores the advantage of specialization and model adaptation to specific subgroups but also highlights the importance of considering demographic factors in the development of classification algorithms.

Given that the dataset comprises posts from the X platform, which are predominantly informal and often contain tags or links irrelevant to model training, we conducted a preprocessing process focused on mentions and links. Specifically, we replaced mentions, represented as "@username," with the generic placeholder "@USER," and links, typically appearing as full URLs, with "HTTPURL." This step was primarily aimed at enhancing the model's performance by removing non-informative elements from the text, thereby facilitating better interpretation and processing of relevant information by the model.

Furthermore, considering that our approach focuses on Spanish-language data, we chose to utilize a pre-trained language model specifically designed for social media texts in Spanish. In this case, we selected the "pysentimiento/robertuito-base-uncased[4]" model available on the Hugging Face platform. This model has been specifically trained to handle colloquial language and the linguistic peculiarities of social media texts in Spanish, making it a suitable tool for our analysis and predictions.

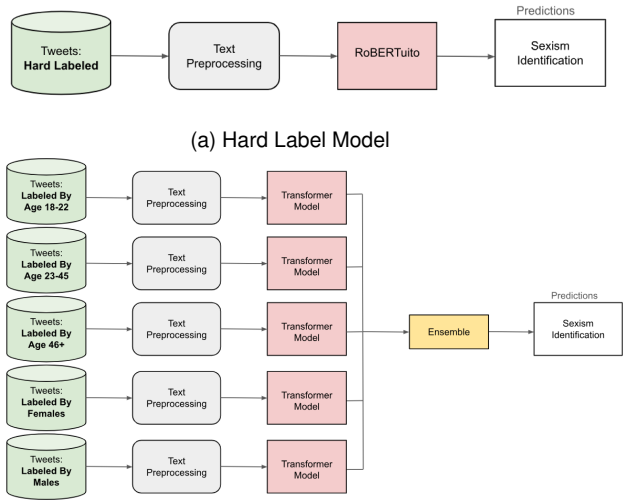
It is important to clarify that during inference, both the "Hard label model" and other classifiers are deterministic, meaning that given the same input data, the model will always produce the same output. However, the fine-tuning or training process might not be deterministic if random seeds are not properly set, which can lead to variations in the results. In this case, the classifiers were run five times, and the resulting F1-scores from these runs were used to create the boxplot. Therefore, the sample size for the boxplot consists of the five F1-scores obtained from these runs, ensuring greater robustness and reliability of the results. Additionally, a hard label was used for all cases in the test set. Data preprocessing and the selection of a language model specialized in Spanish are crucial steps in ensuring that the classification system effectively adapts to the dataset's characteristics and delivers accurate results. By eliminating irrelevant information and utilizing a model optimized for the specific linguistic context, we significantly enhance the model's ability to learn and make more precise predictions based on the actual content of the texts.

Furthermore, the implementation of an "inclusive" ensemble classifier has been shown to be even more beneficial [1]. This classifier combined the results of the group-based models, resulting in superior performance compared to traditional baseline models, particularly, this approach significantly increased recall in the detection of abusive messages, indicating an enhancement in the model's ability to correctly identify this type of content.

These findings emphasize the importance of a more refined and tailored classification strategy, which enhances the overall accuracy in identifying abusive messages. This is especially relevant in the context of content moderation on digital platforms, where the precise and timely identification of abusive language is crucial [1].

To evaluate the effectiveness of this strategy, we started with the hard label model, illustrated

[4] https://huggingface.co/pysentimiento/robertuito-base-uncased



(a) Hard Label Model

(b) Gender and Age Ensemble Model for Sexism Identification

Fig. 1. Models for Sexism Identification

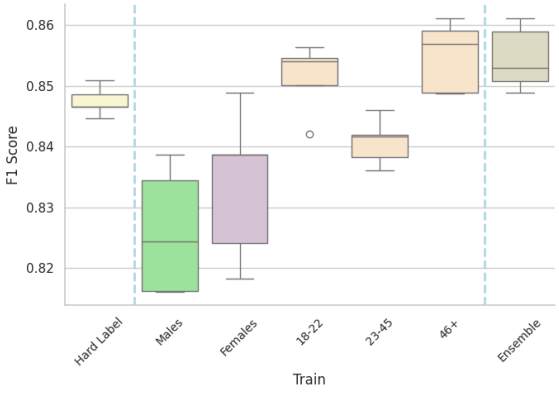


Fig. 2. Comparison Across Different Age and Gender Groups, and Their Ensemble

in Figure 1 (a). Our goal was to assess whether this model could account for multiple perspectives. From this evaluation, we proposed an ensemble-based model for identifying sexism in tweets. This new model integrates the outputs of individual classification models, each specialized in different age and gender groups, as illustrated in Figure 1 (b). By employing a majority voting mechanism, a label is assigned when three or more of these independent models predict the same category.

To ensure the most accurate decision within each ensemble, we use hard voting with a probabilistic voting focus, which relies solely on each model's discrete choice. The probabilities are derived from the softmax output of our transformer model for each prediction. We set a threshold of 0.5 to determine if a prediction is sufficiently confident. For example, if a tweet's label probabilities are [0.50, 0.20, 0.32, 0.43, 0.89], we check each value against the threshold. If the probability of a category is greater than 0.5, it is converted to 1; otherwise, it is set to 0. This process results in a list like [1, 0, 0, 0, 1], indicating that the tweet belongs to category "0" (where category 0 means the tweet is not sexist).

The implementation of this approach not only enhances the robustness of the model by incorporating diverse perspectives from various demographic groups but also optimizes the overall accuracy of the system. By leveraging the specific strengths of each individual model, the model achieves a more precise adaptation to the dataset's characteristics, resulting in improved effectiveness in predicting and classifying labels.

Our model was trained using five different perspectives, considering the viewpoints of both women and men, as well as age groupings (18-22, 23-45, and 46+), as shown in the central section of Figure 2. Focusing on gender groups, we achieved F1 scores of 0.825 ($\sigma = 0.011$) for women and 0.833 ($\sigma = 0.009$) for men. For the age groups, the results were as follows: the 18-22 age group achieved an F1 score of 0.851($\sigma = 0.005$), the 23-45 age group scored 0.840($\sigma = 0.003$), and the 46+ age group scored 0.854($\sigma = 0.005$). By combining these groups using probabilistic voting in an ensemble, we achieved an F1 score of 0.854($\sigma = 0.004$). In addition, the figure presents the results corresponding to a classifier trained with hard labels. It achieved an F1 score of 0.847 ($\sigma = 0.002$).

To assess the statistical significance of these results, we performed an analysis of variance (ANOVA) on the F1 scores for women, men, the hard label, and the ensemble. This analysis revealed a significant difference among these

groups ($p < 0.05$). Specifically, comparing the hard label to the ensemble, and both the male and female groups, as illustrated in Figure 2. Additionally, we performed a further analysis among the age groups and the ensemble. This analysis showed that the groups of 18-22 and 46+, have no statistically significant differences between the performance of the age groups and the overall ensemble ($p > 0.05$). This result suggests that the model performs similarly for the 18-22 and 46+ age groups compared to the overall ensemble, indicating no significant bias in performance across these age groups.

This is important for ensuring fairness, as it shows the model does not disproportionately favor or disadvantage users based on their age in these groups. Nonetheless, the group of 23 to 45 shows a significant difference ($p < 0.05$) compared to the ensemble. In other words, the model's performance in terms of F1 score did not vary significantly between the different age groups and the combined ensemble, suggesting that the ensemble is equally effective in classification, regardless of the considered age group.

These results demonstrate the model's superior performance in most cases, which aligns with the label analysis in Section 2, where the greatest alignment between labels was observed within age groups. The proposed ensemble approach not only maintains consistency across various demographic profiles but also integrates the perspectives of diverse age and gender groups. This inclusiveness enhances the model's coverage and scalability, offering a more comprehensive representation of opinions across different demographic contexts.

4.1 Error Analysis

To better understand how our model makes predictions and where it may go wrong, we analyzed the errors made by our ensemble model. This error analysis is crucial for enhancing the model's ability to distinguish between sexist and non-sexist content, which ultimately improves overall performance. The model also calculates the importance of each word using an attention mechanism, which involves assigning attention scores and normalizing them with the softmax function [23]. In our approach, we applied Hierarchical Attention Networks (HAN) to individual groups—women, men, and age groups (18-22, 23-45, and 46+). Each group produced its own attention scores. We then combined these individual models into a single ensemble model, similar to ensemble methods, to obtain attention values for the consolidated model. This allowed us to perform error analysis on the ensemble model by examining attention scores across all predictions, providing insights into errors such as false positives and false negatives.

Table 5. Confusion Matrix of Predictions

		True Labels	
		Sexism	**No Sexism**
Predicted Labels	**Sexism**	244	155
	No Sexism	17	74

To analyze errors in our ensemble model, we first examine the confusion matrix, which provides insights into the performance of our classifier. The matrix is presented below:

The results indicate a 38.8% rate of false positives and a 18.6% rate of false negatives. False positives, where the model incorrectly labels non-sexist content as sexist, can lead to unwarranted censorship or alienation of benign speech. Conversely, false negatives, where the model fails to identify sexist content, pose a risk of allowing harmful speech to go unaddressed. These errors highlight the critical importance of refining the model to balance sensitivity and specificity, which is further illustrated in the examples and explanations that follow.

To further analyze these errors, we extracted the attention weights assigned by each of the five transformers for each word in the misclassified instances, and these weights were summed and used to color-code the texts, highlighting the most influential words according to their attention scores.

Here are examples illustrating false positives and false negatives:

False Positive: <s>@usuario llevo mi cámara preparada para ser chica de artes q hace fotos a todos <s>

ISSN 2007-9737

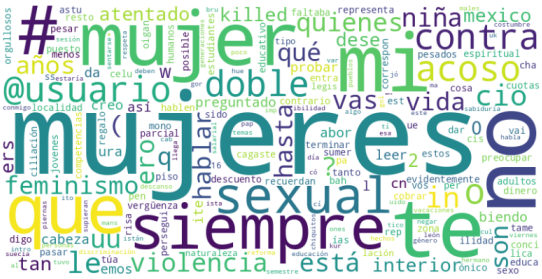


Fig. 3. Word Cloud Representing Common Terms from Tweets Categorized as False Positives in the Ensemble Model for Detecting Sexism

In this example, the model incorrectly classified the statement as sexist. The highlighted phrase "ser chica" (be a girl) might have been flagged by the model as potentially indicative of sexism, especially when considered alongside other terms in the statement. However, while this phrase could be interpreted in various ways, the overall context suggests that the statement is more about taking photos in a social or artistic setting rather than making a sexist comment. The misclassification likely occurred because the model detected a potentially sensitive term but did not sufficiently consider the context, leading to an incorrect label of sexism.

We generated a word cloud illustrated in Figure 3 to visualize the most frequent terms in examples of false positives identified by the model in the context of sexism detection. This word cloud highlights the terms that the model paid the most attention to in each tweet, yet mistakenly classified as sexist. Prominent terms like "woman", "sexual," and "asked" suggest that discussions often revolved around gender-related topics and sexual matters, areas where the model may have misinterpreted context or intent. The frequent appearance of terms related to women and sexuality, such as "woman" and "sexual", points to challenges in distinguishing between neutral or supportive statements about gender and those that genuinely exhibit sexist attitudes. Additionally, terms like "head" and "attack" may reflect contexts where strong or confrontational language was used, leading the model to incorrectly classify the content as sexist. However, it's important to note that the actual meaning depends heavily on the context in which these terms were used. Despite the presence of these words, their interpretation as sexist or non-sexist can vary significantly depending on the surrounding context, which likely contributed to the model's misclassification in these cases. This word cloud serves as a visual tool to understand where the model's predictions might have gone wrong, particularly in discerning nuanced or context-dependent language related to gender issues.

False Negative: <s> @usuario @usuario es que escribes bien... para ser chica <s>

In this case, the model failed to classify the statement as sexist, resulting in a false negative. The phrase "escribes bien... para ser chica" (you write well... for a girl) carries a clear sexist implication by suggesting that writing well is unexpected or noteworthy for someone who is female. This type of backhanded compliment reinforces gender stereotypes by implying that women are generally less capable in this area. The model likely missed this subtle form of sexism because it might not have recognized the underlying bias in the phrase, focusing instead on the more neutral words like "escribes bien" (you write well) without adequately considering the discriminatory context provided by "para ser chica" (for a girl). This oversight led to the incorrect classification as a non-sexist statement.

The model's failure to detect the sexist nature of this statement suggests that it may not have adequately considered the context or the specific combination of words that imply a gender stereotype. This type of error highlights a limitation in the model's ability to recognize subtle forms of sexism, where explicit stereotypes and biases should have been identified. Identifying these terms could be useful for providing feedback to transformers, potentially improving their performance in bias detection.

We generated a word cloud illustrated in Figure 4 to visualize the most frequent terms in examples of false negatives identified by the model in the context of sexism detection. This word cloud highlights the terms that the model paid the most attention to in each tweet, yet still misclassified as non-sexist. Prominent terms

Fig. 4. Word Cloud Representing Common Terms from Tweets Categorized as False Negatives in the Ensemble Model for Detecting Sexism

like "violence," "girl," and "company" suggest that discussions often involved gender-based violence and workplace implications, areas where the model may have struggled to accurately detect subtle or context-dependent sexism. The frequent appearance of terms related to gender and violence, such as "little girl", "girl", and "help", underscores the challenges in detecting potentially sexist language, particularly when it is expressed in nuanced or indirect ways. Additionally, the presence of terms like "ice" and "color" may indicate references to racial or cultural themes that the model also found challenging to classify correctly in the context of sexism.

5 Explaining Model Predictions with ChatGPT

With the aim of facilitating the application of our model in content moderation tasks, we used GPT-4o (2024-08-01) to generate detailed explanations for why certain tweets were labeled as sexist (1) or non-sexist (0). Leveraging a language model like GPT-4o allows for the provision of coherent and understandable interpretations based on the text's context and content, aiding in a better interpretation of the model's decisions. The methodology involves providing ChatGPT-4 with a carefully crafted prompt[5] that outlines the objective of the analysis and requests an explanation for each decision based on the tweet's content and context. This approach enables the generation of detailed responses that justify the classification assigned by our model, offering both a general perspective and specific details that may have influenced the decision.

Below are two examples, one classified as sexist and the other as non-sexist. All five groups within the ensemble have unanimously agreed on these labels.

Sexist Tweet:

"@Diamalware @tuyomiespiritu La mujeres no pueden ser protectoras porque, por regla general, son más débiles físicamente que los hombres. Que algunas "mujeras" se traguen la propaganda globalista demuestra que su coeficiente intelectual está cerca de 80."

(English: @Diamalware @tuyomiespiritu Women cannot be protectors because, as a rule, they are physically weaker than men. That some "womun" buy into globalist propaganda shows that their IQ is close to 80.)

Explanation[6]:

This tweet was classified as sexist because it perpetuates several gender stereotypes. First, it claims that women are inherently physically weaker than men, which is a reductive generalization based on gender. Additionally, it uses the term "mujeras"[7] (a derogatory term) and suggests that women who accept certain ideas have a low IQ. These statements devalue women and reinforce the idea that they are inferior to men in terms of physical and intellectual capabilities, which is a clear form of sexism.

Non-Sexist Tweet:

"@HectorICharles Si piensas que es de gente buena vs gente mala, qué mentalidad tan estrecha tienes."

(English: @HectorICharles If you think it's about good people vs. bad people, you have such a narrow mindset.)

Explanation:

This tweet does not contain sexist elements, as it is simply criticizing a person's perspective or mindset without making any reference to gender or perpetuating

[5]The prompt is available at GitHub

[6]The explanations were originally in Spanish but have been translated.

[7]mujeras[sic]. The correct term in Spanish to refer to women is "mujeres".

gender stereotypes. The criticism is directed at the simplicity of the "good people vs. bad people" thinking and does not imply any judgment or comment based on the sex or gender of the person it is addressed to. Therefore, it is classified as non-sexist.

During the process of using GPT-4o to generate explanations, we encountered several challenges that highlighted the importance of crafting a clear and specific prompt. If the prompt did not explicitly request a detailed explanation of why a particular label was assigned, the model often returned only the label itself or simply confirmed the label's validity without providing additional insights. To address this, we included explicit examples in the prompt to demonstrate how the explanation should be structured for both sexist and non-sexist tweets. Additionally, we noticed that the model tended to skip over explanations for tweets labeled as "non-sexist" unless explicitly instructed to provide them. Another challenge stemmed from the limitations of the free version of ChatGPT-4o, which could not process all the tweets at once, requiring us to split the dataset into smaller parts. Lastly, we observed that ChatGPT consistently treated the provided labels as true without questioning them, which led to the model attempting to justify incorrect labels, such as when a sexist tweet had been misclassified as "non-sexist."

Integrating ChatGPT for generating explanations enhances both the transparency and interpretability of the classification model. By providing clear and contextually informed justifications, this approach allows for a deeper understanding of how and why certain tweets are categorized as sexist or non-sexist. This method not only strengthens the analysis but also serves as a versatile tool that can be adapted to other models and domains.

6 Conclusion and Future Work

The study on sexism detection in the Spanish language, using the EXIST corpus, has revealed significant findings that underscore the complexity of the phenomenon and the need for more nuanced approaches in its analysis. Our research has demonstrated that perceptions of sexism vary considerably among different demographic groups, especially across genders and age groups. Results indicate a 25% discrepancy in the classification of comments as sexist or non-sexist between male and female annotators. This finding highlights substantial differences in sensitivity to sexist content, suggesting that perceptions of sexism are not uniform and may be influenced by gender identity factors.

Analysis of age-related differences also reveals variations, although less pronounced than those observed between genders. Discrepancies among age groups 18-22, 23-45, and 46+ indicate that, while disagreement is less pronounced, there remains diversity in interpreting what constitutes a sexist comment. This finding emphasizes the importance of considering multiple demographic perspectives in developing sexism detection models.

In the context of improving model performance, our focus has been on analyzing error patterns rather than focusing on specific techniques aimed at enhancing accuracy. Error analysis provides valuable insights into the limitations and weaknesses of the current model, allowing us to identify and address underlying issues. By focusing on the types and sources of errors, we can develop more effective strategies to refine the model and improve overall performance. This approach emphasizes understanding and mitigating errors as a key path to enhancement.

The implementation of the combined model revealed a false positive rate of 38.8% and a false negative rate of 18.6%, pinpointing specific areas for improvement in sexism detection. Importantly, the ensemble model outperforms any single classifier trained on data from individual profiles, with the exception of the 46+ profile, where the results are comparable. This underscores the robustness of the ensemble approach and its relevance in ensuring that all perspectives are considered. If hard labels were used to train a single classifier, this would further highlight the ensemble model's ability to integrate diverse viewpoints, making it a valuable tool for addressing the inherent challenges of identifying sexist content.

Incorporating ChatGPT for generating explanations and interpretations has further

enriched our analysis. This tool has provided detailed insights into the features of the texts associated with the model's predictions, offering a clearer understanding of how different tweets are categorized. By leveraging ChatGPT, we have enhanced the model's applicability, making it easier for human content moderators to utilize the model in social media moderation tasks, thus improving transparency and interpretability.

Furthermore, we plan to explore additional modeling techniques, integrating advanced machine learning methods with interpretable models to offer new perspectives on error reduction. We also aim to investigate the impact of regional variations within the Spanish language, as Spanish is a diverse language with many regionalisms. Additionally, exploring factors such as socioeconomic status, educational level, and other user characteristics will provide deeper insights. Implementing these models in real-world and dynamic environments, such as social media platforms, will be crucial for evaluating their effectiveness in evolving data contexts and further validating their performance.

References

1. **Akhtar, S., Basile, V., Patti, V. (2021).** Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection. arXiv preprint arXiv:2106.15896.
2. **Barnes, J., De Clercq, O., Klinger, R. (2023).** Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis.
3. **Blake, K. R., O'Dean, S. M., Lian, J., Denson, T. F. (2021).** Misogynistic tweets correlate with violence against women. Psychological science, Vol. 32, No. 3, pp. 315–325.
4. **Cambridge University Press (2024).** Machismo. https://dictionary.cambridge.org/es/diccionario/ingles/machismo.
5. **Cambridge University Press (2024).** Sexism. https://dictionary.cambridge.org/dictionary/english/ sexism.
6. **Comisión Nacional para Prevenir y Erradicar la Violencia Contra las Mujeres, Gobierno de México (2018).** Frases sexistas que hombres y mujeres debemos dejar de decir para promover la igualdad de género.
7. **de Paula, A. F. M., da Silva, R. F., Schlicht, I. B. (2021).** Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. arXiv preprint arXiv:2111.04551.
8. **Jimenez-Martinez, M. P., Lopez-Nava, I. H., Montes-y Gómez, M. (2024).** An analysis of the impact of gender and age on perceiving and identifying sexist posts. Mexican Conference on Pattern Recognition, Springer, pp. 308–318.
9. **Menczer, F., Fulper, R., Ciampaglia, G. L., Ferrara, E., Ahn, Y., Flammini, A., Lewis, B., Rowe, K. (2015).** Misogynistic language on twitter and sexual violence. Proceedings of the ACM Web Science Workshop on Computational Approaches to Social Modeling (ChASM).
10. **Muti, A., Mancini, E., et al. (2023).** Enriching hate-tuned transformer-based embeddings with emotions for the categorization of sexism. CEUR Workshop Proceedings, CEUR-WS, Vol. 3497, pp. 1012–1023.
11. **Narang, K., Davani, A. M., Mathias, L., Vidgen, B., Talat, Z. (2022).** Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH).
12. **Ojha, A. K., Doğruöz, A. S., Da San Martino, G., Madabushi, H. T., Kumar, R., Sartori, E. (2023).** Proceedings of the 17th international workshop on semantic evaluation (semeval-2023). Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023).
13. **Plaza, L., Carrillo-de Albornoz, J., Morante, R., Amigó, E., Gonzalo, J., Spina, D., Rosso,**

P. (2023). Overview of EXIST 2023: sEXism Identification in Social NeTworks. European Conference on Information Retrieval, Springer, pp. 593–599.

14. **Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V. (2021).** Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation, Vol. 55, pp. 477–523.
15. **Real Academia Española (2024).** Sexismo. https://dle.rae.es/sexismo.
16. **Rodríguez, D. A., Díaz-Ramírez, A., Miranda-Vega, J. E., Trujillo, L., Mejia-Alvarez, P. (2021).** A systematic review of computer science solutions for addressing violence against women and children. IEEE Access, Vol. 9, pp. 114622–114639.
17. **Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Gonzalo, J., Rosso, P., Comet, M., Donoso, T. (2021).** Overview of EXIST 2021: sexism identification in social networks. Procesamiento del Lenguaje Natural, Vol. 67, pp. 195–207.
18. **Rodríguez-Sánchez, F., Carrillo-de Albornoz, J., Plaza, L., Mendieta-Aragón, A., Marco-Remón, G., Makeienko, M., Plaza, M., Gonzalo, J., Spina, D., Rosso, P. (2022).** Overview of EXIST 2022: sexism identification in social networks. Procesamiento del Lenguaje Natural, Vol. 69, pp. 229–240.
19. **Secretaría de Salud, Gobierno de México (2011).** Manual para el uso no sexista del lenguaje.
20. **Tian, L., Huang, N., Zhang, X. (2023).** Efficient multilingual sexism detection via large language models cascades. Working Notes of CLEF.
21. **Vallecillo-Rodríguez, M. E., del Arco, F., Ureña-López, L. A., Martín-Valdivia, M. T., Montejo-Ráez, A. (2023).** Integrating annotator information in transformer fine-tuning for sexism detection. Working Notes of CLEF.
22. **Villa-Cueva, E., Sanchez-Vega, F., López-Monroy, A. P. (2022).** Bi-ensembles of transformer for online bilingual sexism detection. CEUR Workshop Proceedings, Vol. 3202.
23. **Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. (2016).** Hierarchical attention networks for document classification. Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480–1489.

Article received on 15/05/2024; accepted on 20/10/2024.
Corresponding author is Manuel Montes-y-Gómez.